AFRL-OSR-VA-TR-2013-0493

(YIP-10) FAST, FLEXIBLE, RATIONAL INDUCTIVE INFERENCE

THOMAS GRIFFITHS

REGENTS OF THE UNIVERSITY OF CALIFORNIA, THE

08/23/2013
Final Report

DISTRIBUTION A: Distribution approved for public release.

AIR FORCE RESEARCH LABORATORY
AF OFFICE OF SCIENTIFIC RESEARCH (AFOSR)/RSL
ARLINGTON, VIRGINIA 22203
AIR FORCE MATERIEL COMMAND

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE
23 AUG 2013

2. REPORT TYPE

3. DATES COVERED
00-00-2013 to 00-00-2013

4. TITLE AND SUBTITLE
Fast, Flexible, Rational Inductive Inference

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Regents of the University of California,1111 Franklin St.,12th floor,Oakland,CA,94607

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**7**	

FAST, FLEXIBLE, RATIONAL INDUCTIVE INFERENCE

1 Research issues

Despite great advances in artificial intelligence (AI) research over the last fifty years, computers are still far worse than people at solving many important problems, such as learning language, inferring categories of objects from just a few examples, and identifying causal relationships. The goal of this project was to develop automated systems that can match human performance in problems of this kind. The approach that was taken to achieving this goal is one that contributed to the first AI systems: identifying the formal principles that characterize how people solve these problems. This required combining mathematical tools from computer science and statistics with the empirical methods of cognitive psychology. By exploiting the interplay between these disciplines, the resulting research provided insight into how we can make machines learn, and a deeper understanding of how the human mind works.

Part of the challenge of machine learning research is that it requires dealing with a different kind of problem from traditional AI. Learning language, forming categories, and identifying causal relationships are all *inductive* problems, where the limited data available leave the solution radically underdetermined, as opposed to the *deductive* problems such as mathematical reasoning and game-playing in which AI research has traditionally been successful. Induction has a bad reputation in philosophy, having been called a scandal, a riddle, and a myth, and one reason for this is that there is no consensus on exactly how such inferences should be formalized, in contrast to the broadly accepted standard of deductive logic. By studying how people solve such problems, this project aimed to identify some of the formal principles that characterize successful inductive inferences. A general overview of these ideas appears in Tenenbaum, Kemp, Griffiths, and Goodman (2011) and Griffiths, Tenenbaum, and Kemp (2012).

Building on recent work in both AI and cognitive science, this project explored the possibility that Bayesian statistics can provide formal solutions to inductive problems. Bayesian statistics is based upon a simple principle that dictates how a rational agent should change his or her beliefs in light of evidence, called Bayes' rule. Assume that a learner is evaluating a set of hypotheses, and has assigned a "prior" probability $P(h)$ to each hypothesis h in that set. Then, Bayes' rule indicates that after seeing data d, the learner should assign each hypothesis a "posterior" probability $P(h|d)$ proportional to $P(h)$ multiplied by the probability of observing d if h were true, $P(d|h)$. Bayes' rule is a principled way to combine constraints on hypotheses from prior knowledge with the evidence provided by data, and motivates much contemporary research in statistical artificial intelligence and machine learning.

Despite the promise of Bayesian inference as a framework for studying human inductive inference, this approach faces two serious challenges. First, probabilistic inference is extremely computationally intensive, particularly with the large numbers of complex hypotheses needed to model realistic human performance. Second, capturing the kinds of inferences that people are capable of making requires going beyond a simple evaluation of a fixed, discrete set of hypotheses, and considering how we can define models that are flexible, being able to support rich hypothesis spaces that can adapt to accommodate the data. The research supported by this grant focused on addressing these two challenges, with the goal of producing automated systems capable of fast, flexible, rational inductive inference.

The research was divided into two objectives, each addressed by three different lines of work. The remainder of this report summarizes the results of these lines of work. Publications resulting from the work are cited and appear in the reference list. These publications summarize the data collected as part of the grant and describe relevant models at a level where they can be re-implemented.

2 Objective 1: Psychological and neural mechanisms for *fast* inductive inference

Despite their growing popularity in cognitive science, probabilistic models of cognition are often criticized for not identifying psychological or neural mechanisms that might support the computations involved in

Bayesian inference. Part of the issue stems from the fact that probabilistic models provide a different kind of explanation of human cognition than other approaches, answering questions about the abstract computational problems involved in cognition. However, answers at this level of analysis should guide investigation at the levels of algorithm and implementation, and one of the challenges for the probabilistic approach is to develop a clear account of how the underlying computations could be carried out.

Efficient implementation of probabilistic inference is not just a problem in cognitive science – it is an issue that arises in computer science and statistics, resulting in a number of promising solutions. The proposed research will explore whether these solutions seem to correspond to the way that people make inductive inferences, and whether the strategies that people use might lead to new approximation schemes that can be applied in machine learning. In particular, this aspect of the project focused on on the potential implications of Monte Carlo algorithms, which approximate a probability distribution with a set of samples from that distribution. Sophisticated Monte Carlo schemes provide methods for recursively updating a set of samples from a distribution as more data are obtained, providing an answer to the question of how learners with finite memory resources might be able to maintain a distribution over a large hypothesis space.

The research funded by this grant explored three important issues inspired by this idea: how well Monte Carlo methods work as accounts of human inductive inference, how these methods connect to neural computation, and how human cognition might inspire new methods for approximating probabilistic inference. An overview of these ideas appears in Griffiths, Vul, and Sanborn (2012).

Objective 1.1: Evaluating Monte Carlo methods as rational process models

Since no approximation scheme is perfect, we can study whether the way in which different approximations behave correspond to the patterns we see in human behavior. For example, some approximation methods exhibit order effects (particularly primacy effects), something that has been identified as a challenge for Bayesian models of cognition. The research supported by this grant examined whether sophisticated Monte Carlo techniques perform in a way that is similar to human inference across a range of inductive problems.

Our research examined two kinds of Monte Carlo algorithms: particle filters, and Markov chain Monte Carlo. A particle filter approximates a sequentially-updated posterior probability distribution with a set of samples from that distribution, adjusting the samples as more data are observed in order to maintain a good approximation. Abbott and Griffiths (2011) showed that particle filters can produce both primacy and recency effects – two kinds of order effects that emerge in human causal learning. These results demonstrate that this kind of Monte Carlo method might provide a viable explanation for order effects observed for other aspects of human cognition, addressing one of the main empirical challenges to Bayesian models of cognition.

Our work on Markov chain Monte Carlo examined the contexts where a random walk in a representational space might account for human behavior. Abbott, Austerweil, and Griffiths (2013) showed that this simple stochastic mechanism can account for a set of findings in the literature on semantic memory. A random walk on a semantic network reproduces the main empirical phenomena of semantic fluency, demonstrating a "clustering" of semantically related items. This provides a far simpler explanation for these phenomena than previous proposals.

Lieder, Griffiths, and Goodman (2013) took the idea of Markov chain Monte Carlo as a cognitive process in a different direction, examining the question of how long a rational learner should run a simulation before making a decision. The surprising result of this analysis was that when there is a cost for the time spent waiting for an answer, learners should simulate for a fairly short amount of time, resulting in a significant bias in responses. This bias turns out to be the same as that found in experiments on "anchoring and adjustment", providing a compelling rational account of a phenomenon that has typically been taken as evidence of human irrationality.

Objective 1.2: Connecting to neural computation

Monte Carlo methods also provide a way to connect Bayesian inference and artificial neural networks, providing a picture in which existing connectionist approaches provides a substrate for Bayesian inference. Abbott, Hamrick, and Griffiths (2013) explored this connection, demonstrating that one of the Monte Carlo algorithms that has previously been connected to human cognition – importance sampling – can be implemented in a standard neural network model used to capture associative memory. This is an important result, as it is the first demonstration that this kind of probabilistic inference can be performed using distributed representations, potentially providing an avenue for performing probabilistic inference over structured representations in a fixed neural architecture.

In another line of research, Griffiths, Austerweil, and Berthiaume (2012) connected Bayesian inference and neural network models at a more abstract level of analysis. This work showed that simple neural networks can be characterized as performing Bayesian inference with a particular prior distribution, providing a formal foundation for exploring the similarities and differences between these approaches.

Objective 1.3: Developing new approximation algorithms

Studying human cognition can provide clues that lead to new approximation algorithms. For example, human minds often operate under more extreme constraints than most statistical algorithms. In this spirit, we developed a new class of algorithms for probabilistic inference that only maintain a single hypothesis in memory at a time, and only take a single pass through a dataset (Bonawitz, Denison, Chen, Gopnik, & Griffiths, 2011). These algorithms are based on the "win-stay, lose-shift" principle, sticking (stochastically) with a hypothesis as long as it explains observed data, and switching away when it does not. The resulting algorithms can approximate Bayesian inference as well as a single sample from the posterior distribution. A journal article presenting a series of experiments comparing these algorithms to the behavior of adults and children is currently under review (Bonawitz, Denison, Gopnik, & Griffiths, under review).

2.1 Objective 2: Using nonparametric Bayesian methods to define *flexible* probabilistic models

Human inductive inferences are characterized by flexibility: people are able to learn about the world around them without having rigid constraints placed on the hypotheses they consider. For example, we would not assume that there are just three types of things in the world, that objects can only possess a finite number of observable features, or that causal relationships have to be linear. However, this kind of assumption is common in statistical models, where constraints are placed on the number of components in a mixture model, the dimensionality of a latent space, or the functional form of a dependency between two variables.

Recent work in nonparametric Bayesian statistics has begun to make it possible to define probabilistic models that consider an infinite number of structured hypotheses. These models include components that are based on complex stochastic processes, such as the Dirichlet process or Gaussian processes, which define distributions appropriate for nonparametric problems. For example, the Dirichlet process defines a distribution on a discrete but uncountable set of atoms. This makes it useful as a prior in an "infinite" mixture model, where each of these atoms corresponds to the parameters of a mixture component. The resulting model can identify as many clusters as needed in order to capture the structure of observed data, rather than being limited to some finite number a priori.

The research supported by this grant explored connections between nonparametric Bayesian statistics and human cognition, focusing on whether this approach can result in the kind of flexibility that is needed to model human inferences. There were three lines of research: one looked at learning systems of categories, one focused on inferring the features of objects, and one considered how to capture the properties of complex systems involving continuous quantities.

Objective 2.1: Learning systems of categories

Most models of human category learning focus on learning a small number of unrelated categories, but most actual category learning involves learning a large number of related categories. For example, we have to simultaneously learn categories like “dog”, “pet”, and “mammal”. Canini and Griffiths (2011) extended an existing nonparametric Bayesian framework – hierarchical Dirichlet processes – to address learning category structures of this kind, making it possible to infer the structure of a taxonomy relating a set of categories as well as the categories themselves. This approach resulted in a powerful tool for inferring category representations in any taxonomically organized domain, and produced results consistent with human behavior.

Feldman, Griffiths, Goldwater, and Morgan (in press) explored a different way in which categories can be related – through higher-level constraints that govern which categories can appear together. Specifically, when learning phonetic categories – the sounds that make up speech – learning the words that those sounds appear in provides sufficiently strong constraints that it converts the problem from one that is almost impossible to solve into one that can be solved surprisingly well. The resulting system is the first to be able to infer realistic phonetic categories directly from simulated speech data.

Objective 2.2: Forming feature-based representations

Nonparametric Bayesian models can also be used to infer the number and identity of the features that characterize a set of objects. Griffiths and Ghahramani (2011) summarized the key ideas behind a novel approach to solving this problem, based on a stochastic process known as the Indian buffet process. Austerweil and Griffiths (2011) showed that this approach produces results that are consistent with human feature learning, and Austerweil and Griffiths (2010) extended this approach to learn features that are invariant to a particular set of transformations. The results of this work are summarized in a long article forthcoming in *Psychological Review* (Austerweil & Griffiths, in press).

Objective 2.3: Representing objects with continuous dimensions

Developing flexible statistical models that can represent objects with continuous dimensions remains a challenge. The work supported by this grant took on this challenge for two kinds of representations: images and causal relationships. The work on images used Bayesian methods to identify the relevant reference frame and orientation of elements of a scene (Austerweil, Friesen, & Griffiths, 2011), the most representative image in a collection (Abbott, Heller, Ghahramani, & Griffiths, 2011), and the actions that appear in a video stream (Buchsbaum, Canini, & Griffiths, 2011). In each of these cases, a flexible Bayesian model produced predictions that were closely aligned with human performance.

Our work on causal systems considered how existing research on Bayesian models of causal relationships could be extended to capture the continuous elements of these relationships. To this end, we considered how to define appropriate prior distributions on the strength of causal relationships (Yeung & Griffiths, 2011), how people infer causal relationships expressed in continuous time (Pacer & Griffiths, 2012), and how Bayesian inference can be used to explain a set of results in perceptual causality that have traditionally been taken as supporting the use of error-prone heuristics (Sanborn, Mansinghka, & Griffiths, in press). The results of this research are a set of models of human causal reasoning that can directly inform work on causality in machine learning.

3 Impact

The potential impact of the proposed research lies in two areas, corresponding to the two disciplines that will be brought together by this project: psychology and computer science. On the psychology side, the models we have developed provide greater insight into the factors that are involved when people solve inductive problems, allowing us to identify methods for more efficient training and ways of dealing with cases where the approximations that people use lead them to make errors. This work has already been very influential and generated a number of citations – the review article by Tenenbaum et al. (2011) already has 149 citations in Google Scholar, despite having come out in 2011.

On the computer science side, developing a deeper understanding of the formal principles that underlie human inductive inference will pave the way towards building automated systems that match and ultimately exceed human performance. In particular, understanding how people efficiently make inferences over rich (and potentially unbounded) hypothesis spaces will be a first step towards making automated systems that are capable of fast, flexible, rational inductive inference. Our work has already had an impact on computer science, encouraging the development of more sophisticated systems for computer vision that are capable of automatically identifying the features of images.

4 Dissertations supported by this grant

- Canini, K. R. (2011). Nonparametric hierarchical Bayesian models of categorization (Doctoral dissertation). Technical Report UCB/EECS-2011-133, EECS Department, University of California, Berkeley. (Available at: http://www.eecs.berkeley.edu/Pubs/TechRpts/2011/EECS-2011-133.html)

- Austerweil, J. L. (2012). Constructing flexible feature representations using nonparametric Bayesian inference. Unpublished doctoral dissertation. (Available on request.)

5 Personnel supported by this grant (and current positions)

- Kevin Canini (Software Engineer, Google Ventures)

- Joe Austerweil (Assistant Professor, Brown University)

- Christopher Lucas (Lecturer, University of Edinburgh)

- Nick Lewis (graduate student, Indiana University)

- Vincent Berthiaume (postdoctoral researcher, University of Montreal)

- Saiwing Yeung (faculty member, Beijing Institute of Technology)

- Joshua Abbott (graduate student, UC Berkeley)

PUBLICATIONS RESULTING FROM GRANT

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2013). Human memory search as a random walk in a semantic network. In *Advances in Neural Information Processing Systems 25.*

Abbott, J. T., & Griffiths, T. L. (2011). Exploring the influence of particle filter parameters on order effects in causal learning. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society.*

Abbott, J. T., Hamrick, J. B., & Griffiths, T. L. (2013). Approximating Bayesian inference with a sparse distributed memory system. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.*

Abbott, J. T., Heller, K. A., Ghahramani, Z., & Griffiths, T. L. (2011). Testing a Bayesian measure of representativeness using a large image database. In *Advances in Neural Information Processing Systems 24.*

Austerweil, J. L., Friesen, A. L., & Griffiths, T. L. (2011). An ideal observer model for identifying the reference frame of objects. In *Advances in Neural Information Processing Systems 24.*

Austerweil, J. L., & Griffiths, T. L. (2010). Learning invariant features using the transformed Indian buffet process. In *Advances in Neural Information Processing Systems 23.*

Austerweil, J. L., & Griffiths, T. L. (2011). A rational model of the effects of distributional information on feature learning. *Cognitive psychology*, *63*, 173-209.

Austerweil, J. L., & Griffiths, T. L. (in press). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review*.

Bonawitz, L., Denison, S., Chen, A., Gopnik, A., & Griffiths, T. L. (2011). A simple sequential algorithm for approximating Bayesian inference. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.*

Bonawitz, L., Denison, S., Gopnik, A., & Griffiths, T. L. (under review). Win-stay, lose-shift: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology*.

Buchsbaum, D., Canini, K. R., & Griffiths, T. L. (2011). Segmenting and recognizing human action using low-level video features. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.*

Canini, K. R., & Griffiths, T. L. (2011). A nonparametric bayesian model of multi-level category learning. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence.*

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. (in press). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*.

Griffiths, T. L., Austerweil, J. L., & Berthiaume, V. G. (2012). Comparing the inductive biases of simple neural networks and Bayesian models. In *Proceedings of the 34rd Annual Conference of the Cognitive Science Society.*

Griffiths, T. L., & Ghahramani, Z. (2011). The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, *12*, 1185-1224.

Griffiths, T. L., Tenenbaum, J. B., & Kemp, C. (2012). Bayesian inference. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning.* Oxford: Oxford University Press.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*, 263-268.

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2013). Burn-in, bias, and the rationality of anchoring. In *Advances in Neural Information Processing Systems 25.*

Pacer, M., & Griffiths, T. L. (2012). Elements of a rational framework for continuous-time causal induction. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (in press). Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review*.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279-1285.

Yeung, S., & Griffiths, T. L. (2011). Estimating human priors on causal strength. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.*